

C.5.d Graphics Interface.

We have invested a great deal of time and effort in the interface to CONGEN (and GENOA) with the lowest common denominator terminal, a teletype!, in mind. This proved to be a wise decision because many collaborators have only teletypes, or teletype-like hard copy terminals with which to access our programs. However, many structure drawings suffer in teletype form. Some can simply not be laid out within the confines of the teletype grid. In addition, the interface lacks characteristics enjoyed by most chemists, namely the capability actually to draw structural information. Elegant graphics interfaces have been produced for other systems, including Corey's [89] and Wipke's [90] organic synthesis programs, and the NIH Prophet system. In fact, OCONGEN offers capabilities for structure output to a variety of graphics terminals, taking advantage of NIH's Omnigraph package. Two things are clear from this experience. Few potential users can afford currently available terminals, such as the GT-40 series, even though graphical input and output would add considerably to the perceived utility of our programs. It would be out of the question to produce a useful graphics package for a wide variety of terminals, most of which will soon be obsolete.

Yet, we can hardly emphasize three-dimensional aspects of molecular structure and not provide capabilities for visualizing the resulting representations. At Stanford we can utilize the GT-40 purchased under the current grant for some of our work. We do not, however, wish to put a great deal of time and effort into interfaces for that terminal. It is nearly obsolete and compatible, newer versions are simply too expensive for most research groups. We have requested funds to purchase one new graphics terminal in each of the first two years of the proposed grant. The terminal market is changing rapidly with plasma displays, storage displays, and raster displays, potential alternatives to the refresh displays such as the the GT-40. We will seek to purchase a display which is likely to become widely available at relatively low cost and invest our programming in that display.

D SIGNIFICANCE

There are several aspects of our proposed research which we feel are novel and especially significant including not only specific computer programs but also the methods by which they are shared.

One significant aspect of the proposed work, and a primary novelty, lies in the comprehensive computer treatment of both topological and stereochemical aspects of molecular structures in methods to assist scientists in elucidation of important biomolecular structures. By developing these method to include stereochemistry, we will have extended our approaches to computer-assisted structure elucidation to cover the key missing link of determining the actual spatial relationships of atoms in a structure rather than their mere connectivity. Our proposed methods for data interpretation and prediction applied to data collected on unknown structures offer several new treatments of topological representations of structure; the proposed stereochemical efforts, however, are certainly novel because no other similar system for structure elucidation utilizes stereochemistry in comprehensive form. The efforts are also necessary given the strong dependence of spectral properties on molecular configuration and conformation.

The proposed generator of conformers will be a novel solution to a long-standing problem in structural chemistry. Successful completion of a general constrained generator of conformation will be an important result in itself. However, its applications both to candidates for an unknown structure and to studies of conformations of (topologically) known structures will be a more important and novel result. In addition to the proposed collaborative efforts, we foresee several applications to problems relating structure to observed properties, including structure/biological activity relationships. We have not proposed such studies ourselves, but will collaborate with others who can make use of our results, under the resource sharing aspects of our proposal. The investment of resources into developing this program will be repaid many times over by increasing the versatility of CONGEN (as a tool for structure elucidation) and its scope of potential biomedical applications (by providing a link to existing methods based on atomic coordinates).

The GENOA program development, culminating in the SASES system, represents a general, and novel approach, to construction of structures with overlapping substructures, eventually constrained not only by topological but also by stereochemical constraints. These programs will be novel in:

a their modularity, so that they can be used alone or in concert with related programs, of our group or of our collaborators, via shared files of data,

b their comprehensive treatment of stereochemistry,

c the close involvement of the scientist in the problem solving processes of the programs.

The last point perhaps deserves some more emphasis. The most interesting novel result of the proposed work lies in the use to which a well-designed system can be put in the hands of a well-trained chemist. The synergism of man and machine can be a powerful problem solving combination. The structure manipulations embodied in the SASES system represent a "dry" laboratory of functions for manipulating structures and associated data. Many aspects of structural chemistry besides the central task of structure elucidation can be studied utilizing component parts of SASES. In this way, complex questions can be posed and answered on the computer, before execution of actual laboratory experiments.

We feel that these features of our proposed research offer scientists capabilities for solving structures more accurately (in the sense of ensuring that all plausible alternatives have been considered) and in less time. In an era where detection and identification of trace-level organic compounds in environmental and biological milieus is of critical importance, new capabilities such as we propose can be of tremendous value.

We feel that the interest shown by structural chemists in our work, especially the CONGEN program, is already significant. We can point to the workshops, to persons requesting access to the programs at Stanford and to persons interested in the exportable version of CONGEN (see Section F, Collaborative Arrangements, and the Annual Report, Appendix I). However, we have been unable to meet the needs of many of the persons requesting access for the simple reason that methods of access have been limited. Some structural chemists and biochemists, because of personal preference or industrial secrecy, to name only two possible reasons, desire programs in their own laboratories on their own computers. Limitations in laboratory computers and CONGEN's currently limited exportability make export non-trivial. Those who access programs at SUMEX must compete with dozens of other persons for access to the machine and therefore obtain poor interactive service during normal working hours. One significant aspect of our proposal is to promote resource sharing in such a way (the dedicated computer) that export becomes simpler by utilizing exportable languages and more commonly available (and less expensive) computers, while at the same time providing a networked computer environment which greatly facilitates remote access and provides good interaction for those who do not have access to suitable computers in their own institutions.

E FACILITIES AVAILABLE.

This research will be carried out in our well-equipped laboratories in the Department of Chemistry and on the existing SUMEX-AIM computer resource at Stanford, supplemented by the dedicated machine requested in this proposal. We are able to support the proposed work within existing space allocations. Our primary equipment needs, besides the computer, are terminals. We have three character-oriented CRT display terminals, a Tektronix 4012 storage terminal and a GT-40 graphics terminal, plus access to various hard copy terminals when required. The additional terminals requested in the first two years of our grant will provide the required support for visualization of three-dimensional representations of molecular structure, currently possible in very limited ways with the existing Tektronix 4012 and GT-40 display terminals.

The SUMEX computer facilities, which will provide the support for the majority of our development efforts, are shown in Figure 11. This system has a complete complement of peripheral devices to support our needs and to interface with the dedicated system we propose. The system also provides extensive software support which is more than adequate for our proposed program developments, including text editors, a variety of languages, debugging facilities and an excellent staff to assist us in complicated problems related to the SUMEX system or one of its supported pieces of software.

Our personnel will become involved, as they have in the past, with structural problems related to research in my group or in the groups of our collaborators. Many of these problems will require chemical and spectroscopic data to be obtained at Stanford or elsewhere. Any costs associated with collection of these data will be borne by the individual research groups. However, it is important to note that Stanford has well-equipped chemical and spectroscopic laboratories in the Chemistry Department which will enable collection of high quality data in support of these structural studies.

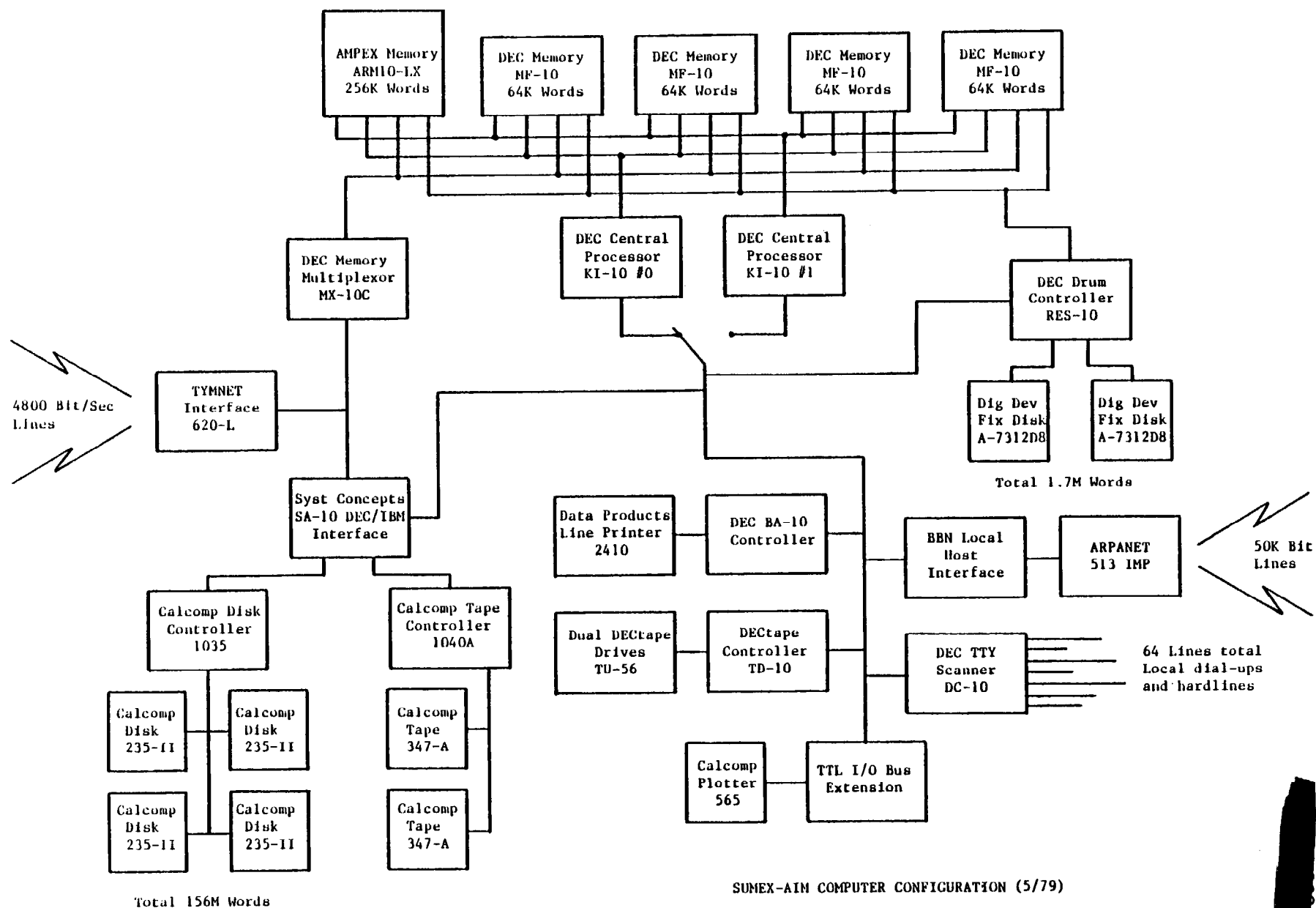


Figure 11. SUMEX hardware

F COLLABORATIVE ARRANGEMENTS.

The proposed improvements to our structure elucidation programs will, in addition, create the possibility of new biomedical applications which make use of existing methods and programs. We propose to promote these new applications via collaborative arrangements.

Collaboration with Prof. David Cowburn.

A very likely application for CONGEN enhanced with a conformation generator would be to the field of conformational analysis. This is the problem of determining the conformation of a structure with known constitution and configuration and is a general problem in describing the structures of molecules. The description of the conformation(s) of molecules of biological origin or of those possessing biological activity is of considerable importance in establishing more clearly the relationship of structure to function in the actions of drugs, hormones, and neurotransmitters on their natural receptors, the mechanism of enzyme action, and the rational design of new drugs. We propose to develop this application by collaboration with Professor David Cowburn and his coworkers at the Rockefeller University in New York. Professor Cowburn is actively engaged in determining peptide conformations using principally nuclear magnetic resonance studies of specifically designed and synthesized isotopic isomers of peptide hormones. These studies use the stable isotopes - deuterium, carbon-13, and nitrogen-15 [91]. Dr. Cowburn now has an account at SUMEX and would use the program remotely, at least at first. It is hoped that an effective collaboration can be developed in which Dr. Cowburn will investigate techniques for effectively rejecting chemically unreasonable conformations as they are generated. Those strategies that may be generally useful will then be adapted for CONGEN and incorporated. These techniques will be related either to general considerations (e.g. insufficient degrees of freedom for cyclization of a particular ring system, from a partially generated conformational state) or to the specific molecules being examined (e.g. restrictions stemming from experimental data such as nmr vicinal coupling constants). Some research using small programs outside CONGEN would be expected to be useful in investigating this area, and some possible techniques are outlined in Dr. Cowburn's letter (pp. -). CONGEN equipped with a conformation generator, would likely be useful to Prof. Cowburn's research in at least three ways:

1) The program would be able to generate all the possible conformations for a given problem with input constraints based on NMR couplings. Such a generation is a difficult task for, e.g., compounds containing large rings. The value of CONGEN would be to provide assurance of exhaustion and to explicitly construct all the possibilities.

2) The program would be able to generate all possible isotopic isomers for a given constitution and configuration. If a pruning technique was available, then the generated list would be extremely useful to Dr. Cowburn in considering the strategies of synthesis and nmr experimentation. The avoidance of particularly costly or time consuming steps is of considerable importance in that experimental work.

3) In conjunction with the spectral interpretation and planning modules proposed, CONGEN may be able to generate strategies for patterns of enrichment or for nmr experiments which are optimum for conformational determination. Some additional programming would probably be necessary to accomplish this.

Collaboration with Prof. Gilda Loew.

Since our proposed conformation generator will output structures with internal (torsional angle) coordinates, it is possible to obtain further information about these structures by doing quantum mechanical energy calculations. By developing a link to these methods, the usefulness of CONGEN should be considerably increased. Since a great deal of work has been done by others on such methods it is not necessary for our group to develop programs of this kind. Instead we propose to develop this link by collaborating with Prof. Gilda Loew and her group in the Dept. of Genetics at Stanford Medical School. Professor Loew's work has involved the use of semi-empirical quantum mechanical energy calculations to derive structure-activity for a variety of drug types [92]. The first step in such a collaboration would be to construct the interface necessary to link the CONGEN output structures with the input for the PCILO (Perturbation Configuration Interaction using Localized Orbitals) program. This program requires as input, structures with internal coordinates. This will be the form of the output from the proposed conformation generator with an assumption of bond lengths and angles. Once this link has been made then we see at least two areas where CONGEN might be helpful to Professor Loew's ongoing research.

1) It will be possible to generate systematically variants of a structure with respect to its constitution, configuration, and conformation. Each such structure would then be given to PCILO for an energy calculation, the results of which are used to help explain potency variations [92]. The advantage of using CONGEN in this way is that an exhaustive generation can be guaranteed which assures no possibilities are overlooked.

2) Professor Loew has been considering the conformational variations caused by the intercalation of ethidium into nucleic acids [93]. The observed stability of such intercalated structures has been related to conformational changes in parts of the DNA structure, in particular, the sugar moieties. The application of CONGEN to such a study would again be a systematic variation of possibilities with particular emphasis on the more difficult cyclic structures.

GUEST Access.


We currently provide access to new users of our programs via the GUEST account at SUMEX. In addition, we have provided GUEST access to several of our recent workshop participants. A list of these participants appears in our annual report (Appendix I). We receive many inquiries about our programs and access to them, far in excess of the current capacity at SUMEX. A list of persons to whom such access has been granted in the past year appears in our annual report (Appendix I). More recent inquiries have been received from (and GUEST access given to):

Prof. Glenn D. Prestwich
Dept. of Chemistry
State University of New York at Stony Brook
Stony Brook, New York 11794

Dr. Andrew Stuper
RCG Group
Rohm and Haas
Norristown and McKean Roads
Springhouse, Pa. 19477

Prof. E. F. Domino
Department of Pharmacology
M6322 Medical Sciences Building
Ann Arbor, Michigan 48109

Djerassi, Carl



Prof. John D. Roberts
Div. of Chem. and Chem. Engr.
Calif. Inst. of Tech.
Pasadena, Calif. 91125



THE ROCKEFELLER UNIVERSITY

1230 YORK AVENUE · NEW YORK, NEW YORK 10021

10 May 1979

Dr. J. Nourse
Department of Chemistry
Stanford University
Stanford, California 94305

Dear Dr. Nourse:

GENERATION OF CONFORMATIONS BY CONGEN

In the past, you and the other members of the Dendral group at Stanford have produced a unique tool for assisting chemists in the interpretation of data, principally in the area of organic structure determination. This tool, the program CONGEN, has recently been augmented by your inclusion of configuration in the descriptive quantities that the program can accept and process.

The possible extension of CONGEN to include conformational values is then the next logical step in the development of this system. In addition to the potential uses of conformer generations in the area of organic structure determination - e.g. inclusion of structural constraints based on experimental observations from NMR or Mass Spectrometric studies, such a program would be of considerable use as a research tool in the applications of conformational analysis in a number of areas. The program would be very useful in exhaustively and precisely defining conformational states in small molecules. In larger molecules (molecular weight >300) the description of standard states (Dunitz 72) and the conformational features associated with certain kinds of rings or regularly repeating substructure are areas of active and significant research endeavor (e.g. [Anet 78], [Dale 75], [Cremer 75]). For polymers, particularly biopolymers, the assumption that a somewhat irregular structure (e.g. a protein) can be described by the cataloguing of standard conformational states associated with the substructures has been widely made ([Desantis 65], [Zimmerman 77]). This assumption has been tested to some degree in a few highly - resolved crystal structures. It has not been tested widely with respect to structures in solution, or to those dynamic structures interconverting relatively rapidly. Such structures, and their analysis, are of very considerable importance in understanding more thoroughly the structure/function relationships of hormones, neurotransmitters and drugs.

It is probable that an enhanced CONGEN could be a very effective tool in these developing fields, for several purposes. For a set of standard states, the program will be able to completely generate all the possible combinations of the conformers of the substructures. This is of considerable importance when a number of rather different substructures are considered in any one molecule, where the permutational task is not straightforward. The powerful features of CONGEN in creating positive and negative conditions for an allowed structure during its generation will provide an exceptionally sound base for testing selections of standard conformational states, for generating a restricted list of possible conformers, and for investigating

Dr. J. Nourse

10 May 1979

conformational classes in sets of related molecules.

The enhanced CONGEN would be useful to our research in all the above ways. I hope that during the period of introducing these enhancements, it will be possible for us to cooperate and collaborate in testing various algorithms that might be incorporated into the final product. These algorithms include various standard modules concerning generation of cartesian coordinates, forced cyclization of coordinates of end groups in rings systems, empirical energy calculations, etc., and modules concerning less well known problems. In this latter area, we are particularly interested in developing techniques for effectively reducing the number of produced conformers during a constrained generation. Methods for doing this may be of a quite general nature. For example, a technique for selection based on the ability of a growing structure to correctly refold to permit cyclization has been described [Dirkx 79]. Exact descriptions for certain regular rings are also possible [Cremer 75]. These pruning techniques may, alternatively, be quite specific to the structure under consideration, e.g. incorporation of known conformational features, or restrictions based on maxima or minima for certain interatomic distances, or restrictions of torsion angles based on predicted effects or specific neighboring groups.

We look forward to being able to pursue this area of research with you, while specifically applying these techniques to the studies of the dynamic conformations of peptide hormones currently under investigation here. The people associated with this research here include Professors William C. Agosta, and David H. Live, Mr. William Wittbold, and myself. Our research in this area is supported by NIH AM-20357.

Sincerely,



David Cowburn
Associate Professor

DC:mmh

- [Anet 78] Anet F. A. L.; Rawdah T. N. "The conformational energy surface of trans,trans,trans-1,5,9-cyclodecatriene." JACS, 1978, 100, 5003-5007.
- [Cremer 75] Cremer D.; Pople J. A. JACS 1975, 97, 1358.
- [Dale 75] Dale J. "Multistep conformational interconversion mechanisms." Topics in stereochemistry, 1975, 9, 199-270.
- [DeSantis 65] DeSantis P.; Giglio E.; Liquori A. M.; Ripamonti A. Nature, 1965, 206, 456-461.
- [Dirkx 79] Dirkx J.; Knappenburg M.; Dufour P. "A program for generation of possible conformations of cyclic molecules." Comp. Prog. Biomed. 1979, 9, 63-68.
- [Dunitz 72] Dunitz J. D.; Waser J. "Geometric constraints in six and eight membered rings." JACS 1972, 94, 5645-5650.
- [Zimmerman 79] Zimmerman S. S.; Pottle M. S.; Nemethy G.; Scheraga H. A. "Conformational analysis of the 20 naturally occurring amino acid residues using ECEPP" Macromolecules 1977, 10, 1-9.

G PRINCIPAL INVESTIGATOR ASSURANCE

The undersigned agrees to accept responsibility for the scientific and technical conduct of the research project and for provision of required progress reports if a grant is awarded as the result of this application.

Date

Principal Investigator

H REFERENCES.

- 1) J. Lederberg,
"DENDRAL-64-A System for Computer Construction,
Enumeration and Notation of Organic Molecules
as Three Structures and Cyclic Graphs,"
(technical reports to NASA)
(1a) Part I. Notational algorithm for tree
structures, 1964, CR.57029
(1b) Part II. Topology of cyclic graphs, 1965, CR.68898
(1c) Part III. Complete chemical graphs; embedding
rings in trees, 1969.
- 2) J. Lederberg, G.L. Sutherland, B.G. Buchanan, E.A. Feigenbaum,
A.V. Robertson, A.M. Duffield, and C. Djerassi,
J.Am.Chem.Soc., 91, 2973, (1969).
- 3) R.E. Carhart, D.H. Smith, H. Brown and C. Djerassi,
J.Am.Chem.Soc., 97, 5755, (1975).
- 4) A.M. Duffield, A.V. Robertson, C. Djerassi, B.G. Buchanan,
G.L. Sutherland, E.A. Feigenbaum, and J. Lederberg,
J.Am.Chem.Soc., 91, 2977, (1969).
- 5) A. Buchs, A.B. Delfino, A.M. Duffield, C. Djerassi, B.G. Buchanan,
E.A. Feigenbaum and J. Lederberg,
Helv.Chim.Acta, 53, 1394, (1970).
- 6) Y.M. Sheikh, A. Buchs, A.B. Delfino, G. Schroll, A.M. Duffield,
C. Djerassi, B.G. Buchanan, G.L. Sutherland, E.A. Feigenbaum,
and J. Lederberg,
Org.Mass Spectrom., 4, 493, (1970).
- 7) (a) L.M. Masinter, N.S. Sridharan, J. Lederberg and D.H. Smith.
J.Amer.Chem.Soc., 96, 7702, (1974).
(b) L.M. Masinter, N.S. Sridharan, R.E. Carhart and D.H. Smith,
ibid, 96, 7714, (1974).
- 8) H. Brown, L. Hjelmeland, and L. Masinter,
Discrete Mathematics, 7, 1, (1974).
- 9) H. Brown and L. Masinter,
Discrete Mathematics, 8, 227, (1974).
- 10) H. Brown,
SIAM Journal of Applied Math, 32, 534, (1977).
- 11) D.H. Smith, B.G. Buchanan, R.S. Engelmores, A.M. Duffield, A. Yeo,
E.A. Feigenbaum, J. Lederberg, and C. Djerassi,
J.Am.Chem.Soc., 94, 5962, (1972).
- 12) D.H. Smith, B.G. Buchanan, W.C. White, E.A. Feigenbaum,
C.Djerassi, and J. Lederberg,
Tetrahedron, 29, 3117, (1973).

- 13) B.G. Buchanan, D.H. Smith, W.C. White, R.J. Gritter,
E.A. Feigenbaum, J. Lederberg, and Carl Djerassi,
J.Am.Chem.Soc., 98, 6168, (1976).
- 14) R.E. Carhart and D.H. Smith,
Computers in Chemistry, 1, 79, (1976).
- 15) T.M. Mitchell and G.M. Schwenger
Organic Magnetic Resonance, 11, 378, (1978).
- 16) G.M. Schwenger and T.M. Mitchell,
in D. Smith, (ed.), Computer Assisted Structure Elucidation,
ACS Symposium Series, 54,
Washington, D.C., 1977, p58.
- 17) (a) C.J. Cheer, D.H. Smith, and C. Djerassi, B. Tursch,
J.C. Braekman, and D. Dalozé,
Tetrahedron, 32, 1807, (1976).
(b) D.H. Smith,
Anal.Chem., 47, 1176, (1975).
(c) D.H. Smith,
J.Chem.Inf.Comp.Sci., 15, 203, (1975).
(d) D.H. Smith, J.P. Konopelski and C. Djerassi,
Org.Mass Spectrom., 11, 86, (1976).
- 18) R.E. Carhart, S.M. Johnson, D.H. Smith, B.G. Buchanan, R.G Dromey,
J. Lederberg,
in P. Lykos (ed.),
Computer Networking and Chemistry,
American Chemistry Society Symposium Series, 19,
Washington, D.C., 1975, pl92.
- 19) T.H. Varkony, R.E. Carhart and D.H. Smith,
in W.T. Wipke and T. Howe, (Eds),
American Chemical Society Symposium Series, 66,
Washington, D.C., 1977, pl88.
- 20) (a) T.H. Varkony, D.H. Smith, and C. Djerassi,
Tetrahedron, 34, 841, (1978).
(b) R.M.K.Carlson, S. Popov, I. Massey, C Delseth, E. Ayanoglu,
T.H. Varkony, and C.Djerassi,
Bioorg.Chem., 7, 453, (1978).
- 21) T.H. Varkony, R.E. Carhart, D.H. Smith, C. Djerassi,
J.Chem.Inf.Comp.Sci., 18, 168, (1978).
- 22) C. Djerassi, D.H. Smith and T.H. Varkony,
Naturwissenschaften, 66, 9 (1979).
- 23) N.A.B. Gray, D.H. Smith, T.H. Varkony, R.E. Carhart,
and B.G. Buchanan.
"Use of a Computer to Identify Unknown Compounds.
The Automation of Scientific Inference,"
Chapter 7 in "Biomedical Applications of Mass Spectrometry,"
G.R. Waller (Ed.),
in press.

- 24) James G. Nourse,
J. Am. Chem. Soc., 101, 1210 (1979)
- 25) James G. Nourse, Raymond E. Carhart, Dennis H. Smith, and
Carl Djerassi,
J. Am. Chem. Soc., 101, 1216 (1979).
- 26) (a) M. Bachiri and G. Mouvier,
Org. Mass Spectrom., 11, 1271, (1976).
(b) G.M. Pesyna and F.W. McLafferty,
Determination of Organic Structures by Physical Methods", Vol 6,
F.C. Nachod, J.J. Zuchermann, and E.W. Randall (Eds),
Academic Press, New York, N.Y., 1976, p91.
- 27) F.W. Mellon,
in "Mass Spectrometry", Vol 4,
R.A.W. Johnstone, Sr. Reporter,
The Chemical Society,
Burlington House, 1977, p89.
- 28) (a) K.S. Kwok, R. Venkataraghavan, and F.W. McLafferty,
J. Am. Chem. Soc., 95, 4185 (1973).
(b) H.E. Dayringer, G.M. Pesyna, R. Venkataraghavan,
and F.W. McLafferty.
Org. Mass Spectrom., 11, 529, (1976).
- 29) S.R. Heller, G.W.A. Milne and R.J. Feldmann,
Science, 195, 253 (1977).
- 30) R.C. Fox,
Anal. Chem., 48, 717 (1976).
- 31) D.L. Dalrymple, C.L. Wilkins, G.W.A. Milne, and S.R. Heller,
Org. Magn. Res., 11, 535, (1978).
- 32) (a) P.R. Naegli and J.T. Clerc,
Anal. Chem., 46, 739a, (1974).
(b) J. Zupan, M. Penca, D. Hadzi and J. Marcel,
Anal. Chem., 49, 2141, (1977).
- 33) D.H. Smith, M. Achenbach, W.J. Yeager, P.J. Anderson, W.L. Fitch,
and T.C. Rindfleisch.
Anal. Chem., 49, 1623, (1977).
- 34) (a) M. Senn, R. Venkataraghavan, and F.W. McLafferty,
J. Am. Chem. Soc., 88, 5593, (1966).
(b) K. Biemann, C. Cone, B.R. Webster and G.P. Arsenault,
J. Am. Chem. Soc., 88, 5598, (1966).
- 35) A. Mandelbaum, P.V. Fennessey and K. Biemann,
Proc. Ann. Conf. Mass Spectrom and Allied Topics, 15th, 111, (1967).
- 36) A. Kundered, R.B. Spencer, and W.L. Budde,
Anal. Chem., <43>, 1086, (1971).

- 37) H.B. Woodruff and M.E. Munk.
Anal.Chim. Acta, 95, 13 (1977).
- 38) H.L. Surprenant and C.N. Reilley,
in "Computer Assisted Structure Elucidation",
D.H. Smith (Ed.),
ACS Symposium Series, 54,
American Chemical Society (1977).
- 39) C.A. Shelley and M.E. Munk,
Anal.Chem., 50, 1522, (1978).
- 40) C.A. Shelley, H.B. Woodruff, C.R. Shelling, and M.E. Munk.
in "Computer Assisted Structure Elucidation",
D.H. Smith (Ed.),
ACS Symposium Series, 54,
American Chemical Society (1977).
- 41) S. Sasaki, Y. Kudo, S. Ochiai and H. Abe,
Mikrochim.Acta, 726 (1971).
- 42) S. Sasaki, H. Abe, Y. Hirota, Y. Ishida, Y. Kudo, S. Ochiai,
K. Saito, and T. Yamasaki,
J.Chem.Inf.Comp.Sci., 18, 211, (1978).
- 43) (a) B.R. Kowalski, P.C. Jurs, T.L. Isenhour, and C.N. Reilley,
Anal.Chem., 41, 1945, (1969).
(b) H.B. Woodruff, G.L. Ritter, S.R. Lowry, and T.L. Isenhour,
Appl.Spectrosc., 30, 213, (1976).
- 44) C.L.Wilkins, R.C. Williams, T.R. Brunner and P.J. McCombie,
J.Am.Chem.Soc., 96, 4182, (1974).
- 45) P.C. Jurs and T.L. Isenhour,
"Chemical Applications of Pattern Recognition",
Wiley-Interscience,
New York, N.Y., (1975).
- 46) J. Schechter and P.C. Jurs,
Appl.Spectrosc., 27, 30 (1973).
- 47) W.E. Brugger, A.J. Stuper, and P.C. Jurs,
J.Chem.Inf.Comp.Sci., 16, 105 (1976).
- 48) (a) M.E. Munk, C.S. Sodano, R.L. McLean, and T.H. Haskell,
J.Am.Chem.Soc., 90, 1087, (1968).
(b) C.A. Shelley, T.R. Hays, M.E. Munk and R.V. Roman,
Anal.Chim. Acta, 103, 121 (1978).
- 49) J.E. Dubois,
in "Computer Representation and Manipulation of
Chemical Information",
W.T. Wipke, S.R. Heller, R.J. Feldmann and E. Hyde, (Eds.),
Wiley-Interscience,
New York, N.Y., 1974, p239.

- 50) L.A. Gribov, M.E. Elyashberg and V.V. Serov,
Anal.Chim.Acta, 95, 97, (1977).
- 51) R. P. Smith,
J. Chem. Phys., 42, 1162 (1965).
- 52) P. J. Flory, U. W. Suter, and M. Mutter,
J. Am. Chem. Soc., 98, 5733, (1976)
and earlier cited references.
- 53) L. E. Scales and J. A. Semlyen,
Polymer, 17, 601, (1976).
- 54) J. B. Hendrickson,
J.Am.Chem.Soc., 86, 4854, (1964).
- 55) M. Dygert, N. Go, H. A. Scheraga,
Macromolecules, 8, 750, (1975).
- 56) J. Dale,
Acta. Chem. Scand., 27, 1115, (1973).
- 57) M. Saunders,
Tetrahedron, 23, 2105, (1967).
- 58) D. F. Bocian, H. M. Pickett, T. C. Rounds, H. L. Strauss,
J.Am.Chem.Soc., 97, 687, (1975).
- 59) J. E. Kilpatrick, K. S. Pitzer, R. Spitzer,
J.Am.Chem.Soc., 69, 2483, (1947).
- 60) D. Cremer and J. A. Pople,
J.Am.Chem.Soc., 97, 1354, (1975).
- 61) C. Altona and M. Sundaralingam,
J.Am.Chem.Soc., 95, 2333, (1975).
- 62) J. Dirkx, M. Knappenburg, P. DuFour,
Comp. Prog. Biomed., 9, 63, (1979).
- 63) A. Murakami and Y. Akahori,
Chem.Pharm. Bull., 25, 2870, (1977).
- 64) C. D. Barry, J. A. Glasel, R. J. P. Williams, A. V. Xavier,
J. Mol. Biol., 84, 471, (1974).
- 65) F. A. Gorin and G. R. Marshall,
Proc.Nat. Acad. Sci., 74, 5179, (1977).
- 66) R.E. Carhart, T.H. Varkony and D.H. Smith,
in "Computer Assisted Structure Elucidation",
D.H. Smith (Ed),
American Chemical Society Symposium Series, 54,
Washington, D.C., 1977, pl26.

- 67) Dennis H. Smith and Raymond E. Carhart,
in "High Performance Mass Spectrometry: Chemical Applications",
M.L. Gross, (Ed.),
American Chemical Society Symposium Series, 70,
Washington, D.C., 1978, p325.
- 68) W. Bremser, M. Klier and E. Meyer,
Org. Magn. Res., 7, 97, (1975).
- 69) W. Bremser,
Fesenius Z.Anal.Chem., 286, 1, (1977).
- 70) W. Bremser,
Anal.Chim. Acta, 103, 355, (1978).
- 71) B.A. Jezl and D.L. Dalrymple,
Anal.Chem., 47, 203, (1975).
- 72) J.T.Clerc and H.Sommerauer,
Anal.Chim. Acta, 95, 33, (1977).
- 73) D.H.Smith and P.C.Jurs,
J.Am.Chem.Soc., 100, 3316, (1978).
- 74) J. Zupan, S.R. Heller, G.W.A. Milne and J.A. Miller,
Anal.Chim. Acta, 103, 141, (1978).
- 75) D.H. Sleeman,
Int.J.Man Machine Studies, 7, 183, (1975).
- 76) G. Beech, R.T. Jones and K. Miller,
Anal.Chem., 46, 714, (1974).
- 77) W. Moffitt, R. B. Woodward, A. Moscovitz, W. Klyne,
and C. Djerassi.
J.Am.Chem.Soc., 83, 4013, (1961).
- 78) J.H. Brewster,
Tetrahedron, 30, 1807, 1974.
- 79) D.N. Kirk and W. Klyne,
J.Chem.Soc. Perkin I, 1076, (1974).
- 80) L. Seamans, A. Moscovitz, G. Barth, E. Bunnenberg, C. Djerassi,
J.Am.Chem.Soc., 94, 6464, (1972).
- 81) R. E. Linder, K. Morrill, J. S. Dixon, G. Barth,
E. Bunnenberg, C. Djerassi, L. Seamans, and A. Moscovitz,
J.Am.Chem.Soc., 99, 727, (1977).
- 82) W. D. Hounshell, D. A. Dougherty, and K. Mislow,
J.Am.Chem.Soc., 100, 3149, (1978).
- 83) A. Kerber,
"Representations of Permutation Groups", Vol II,
Springer-Verlag, N. Y., 1975.

- 84) R.S. Cahn, C. Ingold, and V. Prelog,
Angew.Chem.Int.Ed.Eng, 57, 385 (1966).
- 85) F. A. L. Anet,
Fort. Ch. Forsch., 45, 169, (1974).
- 86) J. H. Dawson, J. R. Trudell, R. E. Linder, G. Barth,
E. Bunnenberg, and C. Djerassi,
Biochemistry, 17, 33, (1978).
- 87) W. C. Johnson,
Ann. Rev. Phys. Chem., 29, 93, (1978).
- 88) C. Marcott, H. A. Havel, J. Overend, A. Moscowitz,
J.Am.Chem.Soc., 100, 7088, (1978).
- 89) E.J. Corey and W.T. Wipke,
Science, 166, 178 (1969).
- 90) W.T. Wipke, H. Braun, G. Smith, F. Choplin, and W. Sielser,
in "Computer Assisted Organic Synthesis",
W.T. Wipke and J. Howe, Eds.,
American Chemical Society Symposium Series, 61,
Washington D.C., 1977, p97.
- 91) (a) D. Cowburn, A.J. Fischman, D.H. Live, W.C. Agosta,
H.R. Wyssbrod,
Proc.Fifth Amer. Peptide Symp.,
Ed. M. Goodman and J. Meinhofer, 322, (1977).
(b) A. J. Fischman, M. Rieman, D. A. Cowburn,
Febs. Letts., 94, 236, (1978).
(c) D. H. Live, and 5 authors ,
J. Am. Chem. Soc., 101, 474, (1979).
- 92) (a) G. Loew and J.R. Jester,
J. Med. Chem., 18, 1051, (1975).
(b) G. Loew, D.S. Berkowitz, and R.C. Newth,
J. Med. Chem., 19, 863, (1976).
(c) G. Loew and R. Sahakian,
J. Med. Chem., 20, 103, (1977).
- 93) G.R. Pack and G. Loew,
Biochim. et Biophys. Acta, 519, 163, (1978).

APPENDIX I

DENDRAL 1978-1979 ANNUAL REPORT

Table of Contents

| Section | Page |
|---|------|
| Subsection | |
| 1. Objectives | 97 |
| 1.1 Overall Objectives | 97 |
| 1.2 Goals for Current Year | 97 |
| 2. STUDIES and RESULTS | 99 |
| 2.1 CONGEN | 99 |
| 2.2 CONGEN Developments | 103 |
| 2.3 RESOURCE SHARING | 107 |
| 2.4 Stereochemistry | 110 |
| 2.5 Structure checking functions for CONGEN | 112 |
| 2.6 Meta-DENDRAL | 121 |
| 2.7 REACT and MAXSUB Programs | 123 |
| 2.8 High Resolution GC/MS System | 124 |
| 2.9 References | 125 |
| 3. SIGNIFICANCE | 127 |
| 4. RESEARCH GOALS 1979-1980 | 128 |

1 Objectives

1.1 Overall Objectives

This progress report covers the second year of our three year grant on computer applications in chemistry, with particular emphasis on techniques of computer-assisted structure elucidation and applications of these techniques to problems of biomolecular structure characterization. To meet this primary objective we have focussed our attention on development of interactive computer programs which are designed to act as chemists' assistants in exploration of the potential structures for unknown compounds. These programs take into account structural information derived from a variety of sources including both physical and chemical methods. We are focussing our research on those aspects of structural analysis which are most difficult to perform manually. We are extending the interpretive power of these programs to enable them to draw meaningful structural conclusions from chemical data. To meet these objectives we are developing a series of computer programs, described in more detail below, which emulate several important aspects of manual approaches to structure elucidation. We are applying these programs to structural problems in our own laboratories and laboratories of others including utilization of the mass spectrometry resource supported under this grant for structural assignment based on mass spectral data.

In order to promote dissemination of the results of our research to the biomedical community, we are developing methods for better access to our programs, including both remote access to the SUMEX resource via nationwide computer networks and exportable versions of important programs including just recently the CONGEN program discussed below.

1.2 Goals for Current Year

Our goals for the current year included the following:

- i) develop and test an exportable version of the CONGEN program and begin investigation of actual export;

- ii) develop CONGEN to utilize techniques of constraint interpretation developed in year 1 and to incorporate other features useful in structure elucidation (see below);

iii) promote resource sharing utilizing SUMEX, through export of programs and through workshops held at Stanford;

iv) interface the stereochemistry package to CONGEN and provide useful output of the STEREO program to

v) the chemist; under Meta-DENDRAL, explore automated rule formation for both mass and ^{13}C spectra, and use such rules to predict spectra and rank candidate structures for an unknown compound based on agreement between predicted and observed spectral properties;

vi) to develop approaches to automated examination of large numbers of candidate structures to aid in experiment planning;

vii) apply the REACT program to investigation of biosynthetic pathways in the marine sterol field;

viii) develop a program for detection of structural similarities in a set of diverse structures;

ix) exploit the high resolution, combined gas chromatography/mass spectrometry system for identification of new natural products.

We have met these goals during the past year, although the methods used and the programs which resulted reflect some changes in emphasis based on requirements of our own applications and those of outside persons using the programs. We have endeavored to place our emphasis on those aspects of the computational methods which were required for certain key structural problems AND which appeared to be of sufficient generality to warrant including in the programs for future use by other persons. Our approaches to experiment planning, use of stereochemistry, definition of aromaticity and mass spectral prediction and ranking all reflect such exposure to real problems and the results represent what we think are generally useful solutions.

2 STUDIES and RESULTS

2.1 CONGEN

2.1.1 Exportable CONGEN

2.1.1.1 Reprogramming CONGEN

Our previous annual report discussed preliminary development of some portions of CONGEN in the BCPL programming language, specifically the structure generation algorithm. This early experience showed BCPL to be a compact and efficient language containing all of the basic features needed for the full reprogramming effort. Continued development has produced a version of CONGEN in BCPL which contains nearly all of the features of the INTERLISP/SAIL/FORTRAN version. The primary exception is the perception of aromaticity, and this feature is currently being implemented. The BCPL version has the following advantages over the previous one;

- a) It requires less than 10% as much computer memory, due partly to the more compact coding and partly to the use of an overlay structure;
- b) It uses about 2-5 times less computer time on typical cases than the most highly-tuned (block compiled) previous version of CONGEN;
- c) The redesigned front-end provides significantly more error checking, a simpler and more flexible input format, and a more thorough "help" facility;
- d) It can easily deal with problems an order of magnitude larger.
- e) It is exportable to a variety of different computers.

2.1.1.2 Overlay structure

As portions of CONGEN were developed in BCPL, estimates of its eventual size could be made and it became obvious that the entire program would occupy a somewhat larger amount of memory than is usually available at many installations (on the order of 100-200 K words, still much smaller than the roughly 450 K words needed by the prior version).

Because the processing in CONGEN falls naturally into several independent activities (generating intermediate structures, imbedding, defining substructures, etc.), the program can easily be broken into separate overlay segments which need to communicate only relatively small amounts of information. In the interest of transportability, though, it was decided not to rely upon the overlay mechanism provided by any particular operating system or language. The safest approach seemed to be to divide the overall program into completely independent, separately runnable modules capable of starting one another and communicating with one another via disk files. The drawback of this approach is that there may be a significant overhead in creating and reading files, and in switching from one module to another. But because all information needed to describe a CONGEN session is maintained on file, the program is unusually robust; even if an error causes the program to crash, CONGEN can simply be restarted and it will restore the complete environment which existed before the erroneous command was issued. Also, a particular operating system may offer some means of accomplishing overlays efficiently, and by interfacing the modules through a small control program, it should be possible to take advantage of such facilities. Under TENEX on the PDP-10, for example, a program may control a large number of forks (independent virtual address spaces) each containing a separate program. We have successfully interfaced the CONGEN modules through a small fork-manipulating program so that the overhead of starting a particular module is paid only once for each CONGEN session.

The current CONGEN is composed of eight modules, the largest of which occupies about 46 K words of memory and the rest of which fall in the range 15-36 K words. We are exploring ways of reducing the size of this largest module (SURVEY - see below) to bring it into this range also. The modules and their functions are as follows:

- a) CONGEN (35 K) - Main control module, user interaction, error checking;
- b) EDITS (19 K) - User interaction for defining substructures;
- c) GENERA (26 K) - Generation of intermediate structures;
- d) PRUNE (15 K) - Elimination of structures based on structural features;
- e) IMBED (36 K) - Expansion of superatoms in intermediate structures;
- f) DRAW (26 K) - Output of structural drawings to the user's terminal;